中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2025)08-2851-15

论文引用格式:Yang X L, Zhang Z C, Xu S K, Zhang B C, Luo X Q and Hu F Y. 2025. Single vertebra 2D/3D registration with fusion of local and global features. Journal of Image and Graphics, 30(8):2851-2865(杨小龙, 张战成, 徐少康, 张宝成, 罗晓清, 胡伏原. 2025. 融合局部与全局特征的单椎体2D/3D配准网络. 中国图象图形学报, 30(8):2851-2865)[DOI:10.11834/jig.240502]

融合局部与全局特征的单椎体2D/3D配准网络

杨小龙[1],张战成[1*],徐少康[2],张宝成[3],罗晓清[4],胡伏原[1]

1. 苏州科技大学电子与信息工程学院,苏州　215009; 2. 上海极睿医疗科技有限公司,上海　200000;
3. 中部战区总医院,武汉　430070; 4. 江南大学人工智能与计算机学院,无锡　214122

摘　要:**目的**　由于患者姿态的变换,术中完整脊柱图像无法与术前CT(computed tomography)形成刚体位置对应,现有的医学图像配准算法在处理脊柱的复杂结构时,常面临配准精度低和鲁棒性不足的问题。针对该问题,提出融合脊柱局部细节特征和全局位置特征的单椎体2D/3D刚性配准网络。**方法**　卷积神经网络通过多组可学习的卷积核增强模型学习椎体的形状和边界等局部结构的能力,Transformer通过自注意力机制能够有效捕捉图像间全局依赖关系并分离出椎体的关键特征,结合两种网络特点,提出双分支网络有效地提取单椎体图像的局部、全局特征。然后设计基于通道、空间注意力的特征融合模块,使网络更好地捕捉椎体信息,并通过多尺度特征逐层优化特征表示,提高网络在不同层次上的感知能力。最后,设计了辅助配准头,利用多层次的空间特征预测姿态参数,使网络在训练过程中逐层优化姿态的预测,从而提高最终的配准精度。**结果**　在Verse数据集上与几种主流的基于迭代优化和基于深度学习的配准方法进行对比实验,本文模型在单椎体配准任务上表现出更高的精度,平均目标配准误差(mean target registration error,mTRE)为1.40 mm,6自由度姿态参数的平均绝对误差(mean absolute error,MAE)为0.008。**结论**　本文提出的配准方法能够获取脊柱局部细节信息以及全局位置信息,从而提高配准精度;且基于多层次特征实现的辅助配准头能够增强监督信息,提高配准模型的稳定性,适用于单椎体的2D/3D医学图像配准任务。源代码可在https://github.com/xlyang2001/Registration获取。

关键词:医学图像;2D/3D配准;单椎体;深度学习;特征融合

Single vertebra 2D/3D registration with fusion of local and global features

Yang Xiaolong[1], Zhang Zhancheng[1*], Xu Shaokang[2], Zhang Baocheng[3], Luo Xiaoqing[4], Hu Fuyuan[1]

1. School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou　215009, China;
2. Shanghai Jirui Medical Technology Co., Ltd., Shanghai　200000, China; 3. Department of Orthopaedics, General Hospital of Central Theater Command, Wuhan　430070, China; 4. Artificial Intelligence and Computer Science, Jiangnan University, Wuxi　214122, China

Abstract: **Objective** Precise pose estimation of intraoperative X-ray images relative to preoperative computed tomography (CT) volumes is a fundamental aspect of 2D/3D registration, which is essential for improving the accuracy of medical analysis and integration in various fields, including surgical planning, radiation therapy, and neural navigation. Traditional methods can be divided into intensity- and feature-based approaches. In intensity-based methods, different ray-tracing techniques are typically used to project 3D volumes for generating simulated 2D X-rays, which are known as digitally recon-

收稿日期:2024-08-27;**修回日期**:2024-12-19;**预印本日期**:2024-12-26

*通信作者:张战成　zczhang@usts.edu.cn

基金项目:国家自然科学基金项目(61772237)

Supported by:National Natural Science Foundation of China(61772237)

structed radiographs (DRR). An optimizer is then employed to find the optimal spatial transformation that maximizes the similarity between the DRR and the corresponding X-ray. Intensity-based methods can achieve high accuracy, but these methods encounter some challenges, such as long registration time due to iterative pose search and the need for generating numerous DRRs for similarity calculation. Moreover, iterative pose search relies on good initialization; otherwise, it may converge to local maximization, which results in registration failure. Meanwhile, feature-based methods typically utilize landmarks for matching to determine the spatial transformation, which can be specific points of anatomical structures or key points. Alternatively, feature detection operators such as Harris, or segmentation can also be employed to extract features. These methods extract features while filtering out a large amount of image information, which leads to higher computation efficiency but lower accuracy than intensity-based methods. Recently, deep learning-based models have emerged as powerful tools for medical image registration. Their feature extraction capability has efficiently addressed the time-consuming issue of traditional methods. The suitable network is designed to learn the feature representation that can describe the complex mapping relationship between images and corresponding labels. This network directly regresses transformation parameters and avoids the need for extensive searching and sampling in the pose space. The intraoperative planar images of complete spines cannot establish rigid correspondence with the preoperative CT, and existing registration algorithms commonly face issues such as low registration and insufficient robustness when dealing with the complete structures of the spine. To address the issue, we proposed a registration method that combines local detail features and global position features for 2D/3D rigid registration in a single vertebra manner. **Method** Convolutional neural networks can enhance the ability of the model to learn the shape, boundaries, and local structures of vertebra, while the Transformer uses the self-attention mechanism to effectively capture global dependencies between images and extract key features of the vertebra. Therefore, by combining the characteristics of both structures, we proposed a multi-stage dual-branch network to effectively extract local and global features from single vertebra images for learning the relationship between features and spatial transformations. The aim was to improve the performance of the regressor. In each stage, the local branch utilizes down-sampling operation and stacked convolution blocks to capture details, edges, and other local information more effectively while reducing computational load. The global branch employs convolution-based patch embedding and multi-head self-attention mechanism to capture the positional relationship between various features and reduce the interference of background in images. The feature fusion module, which is based on the channel and spatial attention mechanism, maps the features of different branches to the same feature space and adaptively fuses local details and global contextual features at different stages, which enhances the expressive ability of the model. Moreover, we progressively optimize feature representations in a coarse-to-fine manner within the network to better capture relevant information. This approach improves the perceptual capability of the network at various scales and levels. Finally, we incorporate auxiliary registration heads to predict pose parameters from the multi-stage fused features, which enhances the supervision information available to network and helps the network gradually optimize pose predictions during training. Ultimately, the final registration accuracy is improved. The input of the network is a single grayscale low-dose X-ray image, and the output is the predicted pose parameters of 6 degrees of freedom (DoFs), which are used to obtain the registered image by the projecting operation. Our network is implemented using the PyTorch framework. The input images are resized to 128 × 128 pixels for training, the learning rate is set to 6e−5, and the weight decay is 0.05. The Adam learning procedure is accelerated using a NVIDIA GeForce RTX 3090 GPU device, which takes approximately 6 h for 320 iterations. **Result** We compared our model with 5 state-of-the-art models on 30 simulated datasets, including 3 iterative optimization-based methods (OPT-GO, OPT-GC, and OPT-NGI) and 2 deep learning-based methods (ResRegNet and EFbackbone-based). The quantitative evaluation metrics contained mean target registration error (mTRE), mean absolute error (MAE), and the registration time, and we provided several visualized registration results of each method. Comparative experiments demonstrated that our model outperforms other methods on the single vertebra datasets. The visualized results also indicate that the registered images have minimal pose deviation from the corresponding target images and demonstrate good alignment, which prove that our proposed method can improve the accuracy in the single vertebra 2D/3D image registration tasks. The average mTRE is 1.40 mm, and the average MAE of 6DoF pose parameters is 0.008. Compared with the second-best method, our model reduces mTRE by approximately 2.70 mm and MAE by 0.02. Furthermore, we conducted a series of ablation experiments and provided corresponding quantitative met-

rics. These experiments clearly demonstrate the effectiveness of each module in our model, including the local branch, global branch, dual-branch feature fusion module, and auxiliary registration heads. **Conclusion** In this study, we proposed a dual-branch single vertebra 2D/3D registration model that contains feature fusion modules to integrate local and global features, which improves the registration accuracy. Moreover, the auxiliary registration heads realized through the features at different layers can enhance supervision information, which increases the stability of the registration model. The experimental results show that our model outperforms several state-of-the-art registration methods and further enhances the registration performance.

Key words: medical image; 2D/3D registration; single vertebra; deep learning; feature fusion

0 引　言

X-ray图像成像快速、易于获取但缺乏三维空间信息，且由于低剂量，图像细节可能不够清晰；而CT（computed tomography）能够清晰显示软组织、骨骼等空间结构，但辐射剂量高。脊柱2D/3D配准是将术中低质2D图像和术前3D CT对齐的一种技术，基于配准关系将3D CT中的详细解剖信息与低质2D成像融合，为医生提供更丰富的术中导航和定位信息（Unberath等，2021），在减少患者辐射暴露的情况下更准确地定位脊柱位置，减少手术创伤，提高手术的效率和成功率（施俊 等，2020）。

传统的方法（Varnavas 等，2015；De Silva 等，2016）可以划分为基于强度和基于特征的方法。在基于强度的配准方法中，通常采用不同的射线追踪方法从3D CT体积中投影生成模拟的2D X-ray，即数字重建射影图像（digital reconstructed radiograph，DRR），然后利用优化器寻找最优的空间变换关系以最大化DRR和X-ray图像之间的相似性（Zhang和Chen，2024）。这些方法可以达到较高的精度但仍存在需要解决的问题，如生成DRR图像涉及大量操作且图像间需要大量的相似性计算（McLaughlin等，2002），难以满足实时性要求；此外，迭代姿态搜索依赖于良好的初始化，否则会收敛到局部极值，导致配准失败。基于特征的方法从图像中提取相应的特征点（Zheng等，2006），如边缘、角点及轮廓等，然后最小化特征之间的距离并通过优化算法找到最优的空间变换。这类方法从图像中提取少量信息，计算量小，效率高（Alam等，2016），但过滤掉大量图像信息，配准算法精度较低。

近年来，基于深度学习的方法已广泛用于2D/3D配准中（Boveiri等，2020；Haskins等，2020）。通过特征学习和隐空间学习代替了相似性测量和优化求解，避免了在姿态空间迭代搜索和采样的需要，能够实现高精度的实时配准且对于噪声、模糊和灰度变换具有较强的鲁棒性。Miao等人（2016）最初利用卷积神经网络（convolutional neural network，CNN）学习图像表征并从中直接回归出姿态参数，实现了实时2D/3D配准。随后，在此基础上提出更多基于CNN的框架（Miao等，2018；Khameneh等，2021）以进一步提高配准的精度。

在临床实时应用中，对完整脊柱进行配准可能会消耗大量的计算资源和时间。此外，脊柱的每节椎骨可能因为不同的病理（如骨折、畸形等）而发生局部变化，外科医生通常只需精确定位和操作特定的椎骨。因此，在单椎体上进行配准能专注于局部变化，从而提供更精确的诊断。

单椎体图像的视野较小，且噪声和伪影可能会模糊重要的椎体边缘等细节信息，对配准精度影响较大，需要细粒度的特征提取以捕获椎体的局部特征，因此考虑利用CNN架构的权重共享和局部性以精确提取椎体的细节特征。但CNN在建模远程依赖关系方面存在局限性，而单椎体图像中存在大量影响特征提取的黑色背景信息，注意力机制能够计算像素之间的相关性以帮助模型关注最相关的区域，因此考虑利用Transformer架构捕捉图像中不同区域的全局关系，去除无关信息并突出椎体区域。为了实现CNN和Transformer两者的优势，已经有了将CNN和Transformer相结合用于图像分割、融合的尝试（Zhang 等，2021；Han 等，2024）。Chen等人（2021a）在U-Net的编码器和解码器之间插入Transformer以建模全局交互，Ma等人（2022）提出一种基于卷积的高效多头自注意力块捕获局部空间上下文信息，Mok和Chung（2022）提出一个由粗到细的视觉变换器用于仿射配准，对patch嵌入和前馈层进行改

进以补充局部性，但这些工作集中在CNN和Transformer两个架构之间的替换和顺序堆叠，会导致信息的丢失且特征融合不够充分。

因此，本文考虑使用一种双分支结构在不同尺度下和层次下将CNN提取到的椎体细节特征和Transformer捕捉的图像边界、背景等上下文信息进行相互补充，形成更完整的特征表示，帮助网络理解椎体在整个图像中的位置和姿态，以改善单椎体2D/3D配准的精度。主要工作有：1）提出一个双分支局部与全局特征融合网络LGFF-Reg（local and global feature fusion network for registration）用于单椎体2D/3D图像配准。局部分支使用卷积块捕获单椎体的结构细节信息；全局分支利用多头自注意力块获取椎体的整体位置信息，并设计双分支特征融合（dual branch feature fusion，DBFF）模块进行信息融合，学习特征与空间变换之间的映射关系，从而提高回归器的性能。2）设计了基于通道和空间注意力的双分支特征融合模块，将双分支特征映射到相同的特征空间并在不同尺度下自适应地融合局部细节和全局上下文特征，从而提升特征表达能力。3）设计了利用网络多层次特征分别预测姿态参数的辅助配准头，并计算其预测值与真实空间变换参数之间的损失，从而增强网络的监督信息，进一步提高配准精度。

1 相关工作

1.1 基于CNN的2D/3D配准方法

基于深度学习的方法已经证明了其在配准方面的精度和实时性能（Chen等，2021b）。Miao等人（2016）提出基于分层学习的姿态估计方法，利用CNN实现具有大捕获范围和高精度的实时2D/3D配准，训练回归器来恢复DRR和X-ray图像到其基本变换参数之间的差异并直接预测配准变换参数。但是，其使用的是相对简单、浅层的网络结构，难以处理有复杂结构的图像且模型的拟合能力能够进一步提升。在此基础上，Meng等人（2022）设计基于多个残差块的网络学习图像特征并回归出目标和初始变换参数之间的6自由度（degree of freedom，DoF）差异。Zhang等人（2023）采用正则化自编码器和多头自注意力机制设计患者特定的自监督模型预测X-ray源的6DoF位姿，并获取相对于姿态参数的梯度以进一步改进预测的姿态。徐少康等人（2023）提出一种基于自编码器的姿态回归网络，通过隐空间解码捕获几何姿态信息，并利用基于梯度的归一化互相关作为损失进行“细配”以保证姿态回归的精度。Gao等人（2024）设计了一个使用投影空间变换器的完全可微的框架，并利用双后向梯度驱动损失函数去进行训练以扩展2D/3D配准的捕获范围。Sun等人（2024）提出一种基于透视投影三角性特征的快速X-ray/CT图像配准方法，利用点特征描述符构建具有旋转、平移和尺度不变性的特征实现6个变换参数的解耦，但患者解剖结构的变化和图像噪声等会影响特征的提取和配准过程。为了进一步提高在真实临床数据集上的配准精度，Aubert等人（2023）在配准之前使用基于生成对抗网络（generative adversarial network，GAN）的跨模态图像到图像变换过程将DRR图像转换为X-ray图像以减少对相似性度量的选择；而Zheng等人（2022）在源域（DRR）上训练模型，建立外观—姿态关系后再使用无监督跨域自适应网络使模型适应目标域（X-ray），王熙源等人（2024）将源域上捕获的图像特征与空间变换间的对应关系迁移到目标域，并借助公共特征减少域间特征差异。这些方法表明CNN能够从输入图像中学习到不同患者、不同噪声水平下的特征，直接预测姿态参数，展现了在配准精度、实时性上的优势。

1.2 基于Transformer的医学图像配准

CNN架构具有权重共享和局部性的归纳偏差，但在建模显式远程依赖关系方面存在局限性。而视觉变换器（vision Transformer，ViT）利用自注意力机制在非重叠的图像patch序列中捕获图像中的远程依赖关系，这一特性使得ViT在许多医学成像应用中展现了先进的性能。Mok和Chung（2022）提出一种由粗到细的视觉变换器用于配准，通过自注意力算子的全局连通性以及卷积前馈层的局部性，将图像对的全局方向、空间位置和长期依赖关系编码为一组几何变换参数。Wang等人（2024）设计了包含Transformer的残差连接增强长序列映射能力，并利用对比学习策略自适应地在全局范围内检索有效特征，保证在遮挡情况下的脊柱2D/3D配准性能。Lan等人（2023）提出基于可变形区域的结构相关嵌入模块，这种方法能够自适应地将图像分割成具有方向约束的不同大小和形状的可变形结构区域，并有效地保持区域之间的语义一致性，从而计算出精确的

变换关系。Chen 等人(2024)提出基于 Transformer 的多层双流特征匹配网络用于配准，该网络在每个分支独立进行特征提取，利用自注意力机制的查询—匹配思想实现图像对之间的显式特征匹配，并学习更新后的特征到变形场的映射。这些方法表明 ViT 能够更好地理解图像之间的空间对应关系，展现了其在医学图像配准任务中的有效性。

1.3 单椎体配准

脊柱是一个复杂的周期性结构，在不同患者之间差异很大，且环境中相似的相邻结构和重叠的存在会影响配准的性能。因此，Gill 等人(2012)提出一种基于生物力学的配准方法，将 CT 上的每个腰椎作为一个子体积分别进行转换，Hille 等人(2018)将全局非刚性配准分解成多个局部刚性配准，以准确模拟建模不同患者移动引起的脊柱变形问题。Aubert 等人(2023)使用跨模态图像到图像转换模型将 X-ray 图像转换为一组骨骼分离的类似 DRR 图像作为 3D/2D 配准前的先验步骤，以避免相似结构的不匹配，且只在感兴趣的孤立结构上(即每个椎体)测量相似性水平，减少了对相似性度量选择的依赖并提高了椎骨精细配准的效果。为了评估疾病进展和手术结果，通常进行纵向脊柱配准，Sanhinova 等人(2024)利用深度学习分割模型预测分割掩码，然后在椎体水平下利用获得的分割掩码将后续对象配准到基线。

刚性配准可以防止骨结构变形，而节段级配准可以确保每个椎体具有独特的变换矩阵。因此，相较于在完整脊柱上进行配准，单椎体配准能够更准确地对比特定椎体的变化并进行独立处理，从而避免误差累积，确保更高的精度。

2 方 法

2.1 2D/3D 配准问题描述

为了进行 2D/3D 配准，需要将两种不同维度的图像统一到同一空间维度，通常将 3D 体数据投影到与待配准图像维度一致的平面，使得问题转换为求解当前 2D 图像相似度达到最值时的空间投影变换问题，可描述为

$$\boldsymbol{T}^{*}=\underset{T}{\operatorname{argmax}}\, S(\boldsymbol{I}_F, \boldsymbol{P}(\boldsymbol{T}(\boldsymbol{V}_M))) \tag{1}$$

式中，$\boldsymbol{V}_M$表示 3D 浮动体积(CT)，$\boldsymbol{I}_F$表示 2D 固定图像(X-ray)，函数 S 表示固定图像与投影图像之间的相似性度量，$\boldsymbol{T}$和$\boldsymbol{T}^{*}$分别表示配准前后的空间变换。$\boldsymbol{P}$表示由 CT 体数据到 DRR 图像的投影，仅与成像仪器有关，且当术中 2D 图像的成像参数固定时，投影$\boldsymbol{P}$也是唯一确定的。配准后的图像$\boldsymbol{I}_R$可表示为

$$\boldsymbol{I}_R=\boldsymbol{P}(\boldsymbol{T}^{*}(\boldsymbol{V}_M)) \tag{2}$$

式中，$\boldsymbol{T}^{*}$为最优空间变换。

2D/3D 刚性配准中，患者术中姿态和 CT 之间的对准可由 $\boldsymbol{T}$表示，将 CT 从初始位置变换到同一坐标系下的患者术中位姿。$\boldsymbol{T}$由 3 次平移$\boldsymbol{t}=(t_x, t_y, t_z)$和 3 次绕轴的旋转$\boldsymbol{r}=(r_x, r_y, r_z)$参数化，可表示为

$$\boldsymbol{T}=\begin{pmatrix}\boldsymbol{R}(\boldsymbol{r}) & \boldsymbol{t}\\ 0 & 1\end{pmatrix} \tag{3}$$

式中，$\boldsymbol{R}$是控制 CT 体积围绕原点旋转的旋转矩阵。

在基于深度学习的方法中，通常将 2D/3D 配准表述为一个回归问题，在以姿态参数作为标签的监督学习方法下，通过设计回归器学习图像特征与空间变换之间的复杂非线性关系，从而直接回归出变换参数。在训练过程中，模型目标是最小化预测的变换参数与真实标签之间的差异，从而使得利用该变换参数对 CT 投影得到的图像与 X-ray 图像相似性最大，最终实现 2D/3D 配准。

2.2 配准模型结构

2D/3D 配准问题可转换为从 2D X-ray 图像中寻找最佳 CT 投影姿态的问题。本文配准网络整体结构如图 1(a)所示，其中网络的输入为单幅灰度 X-ray 图像，输出为预测的 6DoF 姿态参数，Patch 嵌入以及下采样模块的大小和高度分别表示网络当前层次特征图的尺寸以及通道数。图 1(b)描述了相关子模块的详细结构。

首先，通过一个由两个堆叠的 3×3 卷积、批归一化(batch normalization，BN)、激活函数(rectified linear unit，ReLU)操作构成的 Stem Block 提取初始局部特征，保证输出的特征图具有统一的大小和特征表示，然后将其提供给双分支网络。局部分支利用堆叠的下采样和卷积块实现多尺度的局部细节特征提取，而全局分支利用 patch 嵌入和多头自注意力块提取单椎体 X-ray 图像的全局上下文信息及位置信息，构建不同尺度的特征图，使模型能够更好地捕捉多层次信息；然后使用提出的 DBFF 模块在不同层次自适应地融合双分支的特征，并利用辅助配准头学习更有效的特征表示；最后通过卷积和全连接

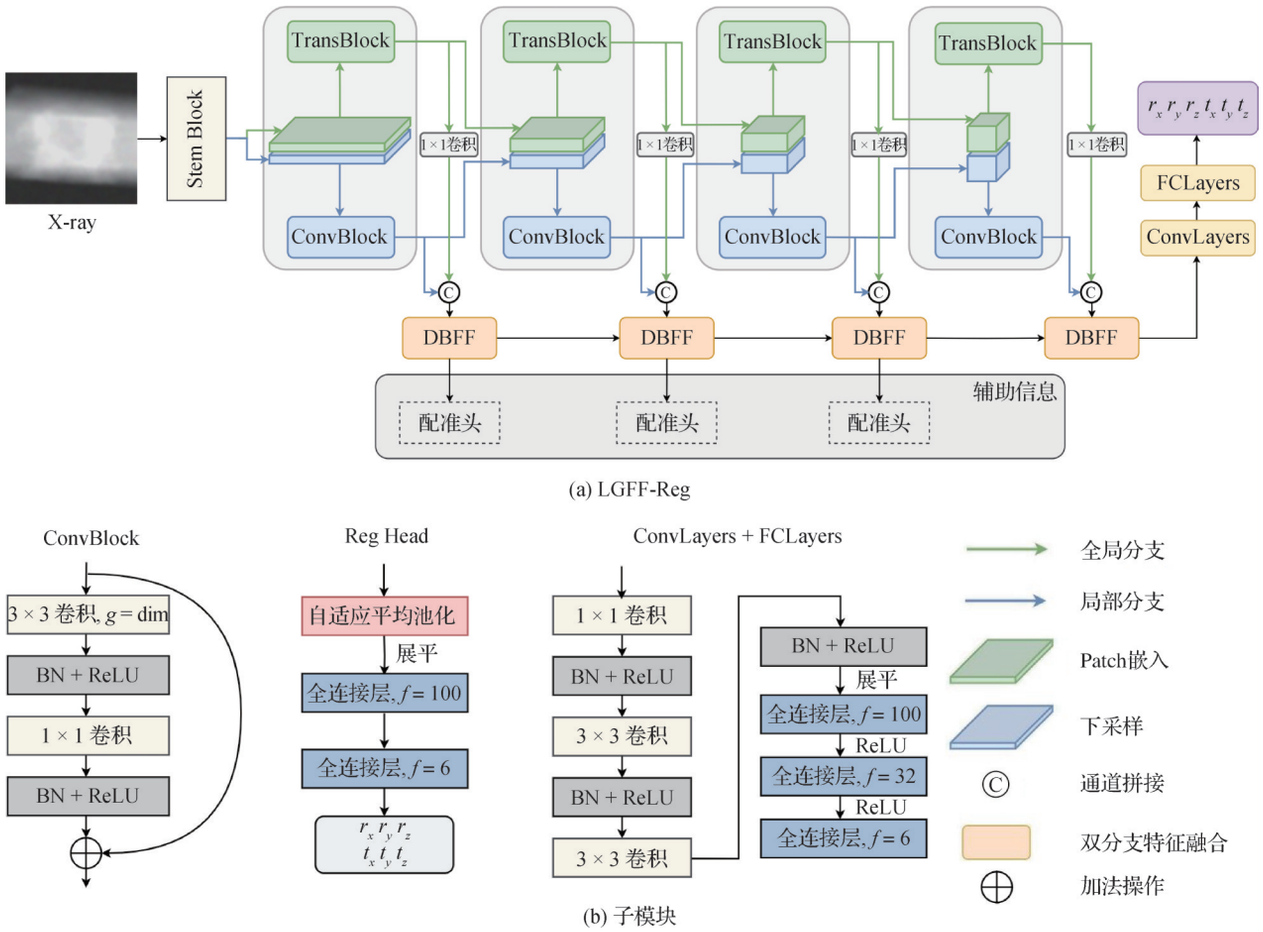

图1　配准网络整体结构

Fig. 1　Overall architecture of the registration network ((a) LGFF-Reg; (b) sub module)

操作从融合的特征中推断出6 DoF姿态参数。

2.2.1　局部分支

局部分支(local branch, LB)包括4个层次,在不同尺度上进行卷积操作,生成一系列尺寸逐层减小、通道数逐层增加的特征图,以由粗到细的方式提取图像的局部特征。每个层次包括一个步长为2的3×3卷积实现图像的下采样,以及两个堆叠的深度卷积块组成的ConvBlock。深度卷积块由深度卷积和点卷积组成并带有BN、ReLU和残差连接。深度可分离卷积核大小设置为3,分组数与当前层次下通道数一致,能够减小计算量和参数量;点卷积通过1×1卷积实现,主要用于调整通道数并整合通道信息。深度卷积在空间维度上进行卷积,而点卷积在通道维度上进行卷积,组合后能够在减少计算量的同时更有效地捕捉单椎体图像中的细节和边缘等局部信息。

2.2.2　全局分支

CNN在提取局部特征方面效果显著,但在捕捉全局特征和处理长距离依赖关系时存在局限性。在单椎体配准中,获取全局信息有助于模型更好地定位脊柱位置,并减少对无关背景信息的注意,从而提高配准精度。为此,本文设计了全局分支(global branch, GB),利用Transformer结构中的自注意力(self-attention, SA)机制来捕捉图像中长距离依赖关系。通过多个注意力头在不同子空间内学习特征,全局分支能够有效捕获不同尺度的信息并实现更丰富的特征表达。

在全局分支中,每个层次的特征提取模块包括一个Patch嵌入和一个TransBlock块。Patch嵌入通过卷积层实现,卷积层的步长与卷积核大小一致,因此卷积结果相当于对图像进行下采样,再进一步将特征图映射到高维向量空间;TransBlock可划分为多头自注意力(multi-head self-attention, MHSA)、层归一化(layer normalization, LN)和多层感知机(multilayer perceptron, MLP)部分,具体结构如图2所示。

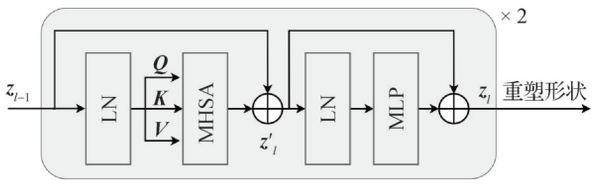

图2　TransBlock结构

Fig. 2　Architecture of TransBlock

MLP是一个由两层全连接构成的前馈层，用于增强网络的表达能力。注意力机制的基本原理是将输入特征通过线性变换得到查询(query，$\boldsymbol{Q}$)、键(key，$\boldsymbol{K}$)和值(value，$\boldsymbol{V}$)矩阵，计算$\boldsymbol{Q}$和$\boldsymbol{K}$之间的相似度并归一化为权重，然后对值进行加权求和，具体为

$$f_{\text{Attention}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=f_{\text{softmax}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{4}$$

式中，d表示键$\boldsymbol{K}$的维度。输入z_{l-1}经过MHSA、LN和残差连接模块后输出为z'_l，再经过MLP、LN和残差后输出为z_l，其中l表示当前Block，整体的前向计算过程为

$$\begin{aligned} z'_l &= MHSA(LN(z_{l-1}))+z_{l-1} \\ z_l &= MLP(LN(z'_l))+z'_l \end{aligned} \tag{5}$$

单椎体图像中存在大量非目标区域，这些背景信息可能会干扰配准过程，因此利用全局分支通过上下文信息更好地区分椎体结构和背景信息，从而提高单椎体特征提取的准确性。

2.2.3　双分支特征融合模块

为了更有效地融合局部分支和全局分支的特征，本文设计了一个DBFF模块来融合不同层次的局部特征和全局表征，并连接网络前一层融合后的语义信息。由于通道注意力(channel attention，CA)能够捕捉通道间的相互依赖性，并提升特定语义特征的表示能力，而空间注意力(spatial attention，SA)机制能够增强局部细节并抑制无关区域，两者结合，模型可以更精确地定位和提取单椎体的结构和位置信息。DBFF模块的详细结构如图3所示。

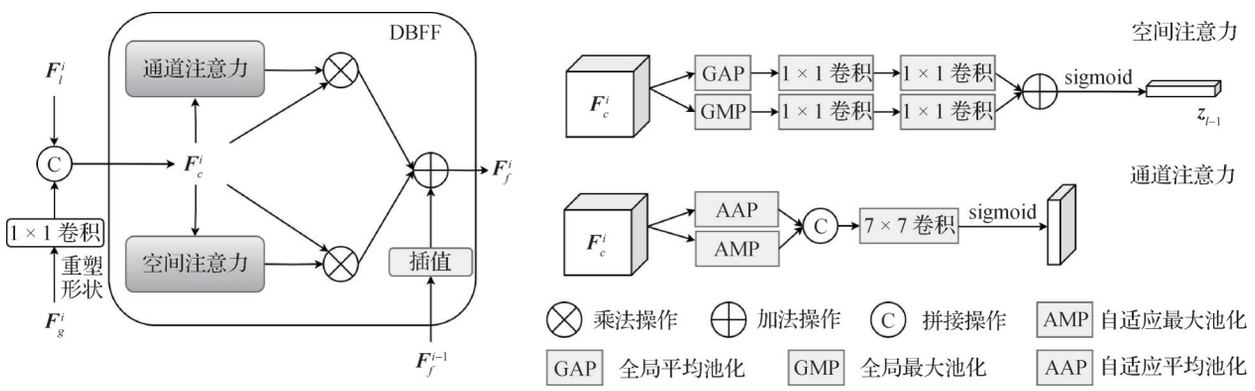

图3　DBFF模块结构

Fig. 3　Architecture of the DBFF module

本文利用下采样实现多尺度的特征提取，形成了一系列不同大小的特征图，并在不同层次上逐步融合双分支特征。在网络i层，局部分支得到的特征图$\boldsymbol{F}_l^i$维度为$C\times H\times W$。其中，C、H、W分别为通道、高度和宽度。而全局分支得到的patch嵌入维度为$K\times E$。其中K，E分别表示图像的patch数量和嵌入维度。通过转置和重塑操作后得到特征图$\boldsymbol{F}_g^i$，然后使用1 × 1卷积调整通道数并将特征向量投影到另一个特征空间。最后，在通道上进行拼接得到特征图$\boldsymbol{F}_c^i$。双分支特征融合的具体过程可表示为

$$\begin{aligned} \boldsymbol{F}_f^i &= f_{\text{Interp}}\left(\boldsymbol{F}_{f-1}^i\right)+\boldsymbol{F}_c^i\otimes\left[CA\left(\boldsymbol{F}_c^i\right)+SA\left(\boldsymbol{F}_c^i\right)\right] \\ \boldsymbol{F}_c^i &= f_{\text{Concat}}\left(\boldsymbol{F}_l^i,f^{1\times1}\left(\boldsymbol{F}_g^i\right)\right) \end{aligned} \tag{6}$$

式中，$f_{\text{Interp}}(\cdot)$表示双线性插值操作，以保证特征融合时特征图尺寸一致；⊗表示按元素相乘；$f_{\text{Concat}}(\cdot,\cdot)$表示在通道上进行拼接操作，$f^{1\times1}$表示卷积核大小为1的卷积操作。$CA$和$SA$分别表示通道注意力和空间注意力操作，可表示为

$$\begin{aligned} CA(x) &= \sigma\left(MLP\left(GAP(x)\right)+MLP\left(GMP(x)\right)\right) \\ SA(x) &= \sigma\left(f^{7\times7}\left(f_{\text{Concat}}\left(AAP(x),AMP(x)\right)\right)\right) \end{aligned} \tag{7}$$

式中，GAP、GMP分别表示全局平均池化和全局最大池化，AAP、AMP分别表示自适应平均池化和自适应最大池化，σ表示sigmoid函数。

通过这种融合策略，模型能够在多层次上集成局部和全局信息，表达单椎体配准中的微小位姿的

变化，学习隐空间特征与空间位姿之间的复杂非线性关系，从而提升配准的精度和鲁棒性。

2.2.4　辅助配准头

通过多层下采样操作生成的多尺度特征，网络能够综合不同尺度的信息，增强对图像变化的适应能力。为了进一步提升配准精度，在网络不同层次上添加辅助配准头（reg head），通过池化操作对特征图进行压缩，并利用全连接层将特征映射到固定大小的输出向量。在计算量增加较小的情况对特征图进行处理并输出6个姿态参数，可以集成到配准网络中，并在每层计算姿态参数与标签之间的损失，能够为模型提供细粒度的监督信号，从而提升模型的整体特征表示能力。

在训练阶段，辅助配准头的目的是通过不同层次姿态预测值的损失引导网络在中间层有效的特征表示，并加速训练过程的收敛速度。

2.3　损失函数

本文利用网络预测值和标签值之间的平均绝对误差（mean absolute error，MAE）作为损失函数监督网络训练，具体为

$$L_{\mathrm{MAE}}=\frac{1}{N}\sum_{i=1}^{N}\left|P_g^i-P_p^i\right| \tag{8}$$

式中，N是样本总数，P表示6DoF姿态参数$\left(r_x,r_y,r_z,t_x,t_y,t_z\right)$，下标$g$和$p$分别表示真实值和预测值。

辅助配准头回归出的姿态参数都和标签计算损失以增强网络的监督信息，帮助网络更好地学习特征与姿态变换参数之间的映射关系。因此，在训练过程中，损失函数可表示为

$$L_{\mathrm{total}}=\sum_{i=1}^{3}\lambda_i\times L_{\mathrm{aux}_i}+L_{\mathrm{main}} \tag{9}$$

式中，λ_i表示权重因子，用于控制辅助损失的影响。L_{aux_i}表示第i个辅助配准头的损失，L_{main}表示网络输出的损失，二者都表示与姿态参数之间的MAE损失。

3　实　验

在临床应用中，获得大量真实的术中X-ray以及相应的姿态参数进行网络训练是困难的。因此，本文利用预处理后的DRR图像为网络输入，而相应的投影参数被保存作为标签，最终回归出6DoF姿态参数。

3.1　数据集及预处理

不同患者的脊柱在形状、大小等方面存在差异，图像间存在偏差，因此配准过程需要考虑患者的个体差异，这是一个患者特定的过程。本文使用的数据集来自Verse分割数据集（Löffler等，2020），其中包含脊柱CT图像的分割掩码。在脊柱手术过程中，患者的体态可能导致椎体之间的位置发生变化。因此为了减少由于椎体移动和变形引起误差导致术中完整脊柱图像无法与术前CT形成刚体位置的情况，外科医生通常会采用单椎体配准的方式。

本文选择了30个具有清晰脊柱结构的CT，首先按照每3个椎骨为一组进行分割，并裁剪出相应区域，将其归一化为128×128×128的尺寸，并进一步裁剪出单节椎体用于投影生成相应的DRR图像，给定变换矩阵$\boldsymbol{T}$和CT体积$\boldsymbol{V}$，DRR图像生成过程可表示为

$$\boldsymbol{I}_{\mathrm{DRR}}(x)=\int_{p\in l(x)}\boldsymbol{V}\left(\boldsymbol{T}^{-1}p\right)\mathrm{d}p \tag{10}$$

式中，$l(x)$表示连接X-ray源和图像平面上点x的线积分，p表示$l(x)$上的点，且由成像模型确定。

实验中，DRR图像大小为128×128像素，旋转和平移变换参数的变换范围分别设置为[−10 mm，10 mm]和[−18°，18°]。通过在该范围内均匀采样，对每个CT随机投影生成10 000幅DRR，其中9 000幅用于训练，剩余1 000幅用于测试。在DRR图像中，脊柱通常显示出清晰的细节和轮廓信息，而临床中的低剂量X-ray通常成像质量较低。为了减少图像差异对配准性能的影响，在DRR中添加雾化、模糊和伪影并调节对比度和亮度，以模拟真实手术环境中的低剂量X-ray情况，数据集合成的过程如图4所示。

数据集合成的具体流程如下：首先创建与输入图像大小相同的白色叠加层并生成随机噪声，按照给定的雾化强度进行加权混合生成雾化效果，与原始DRR图像进行加权混合得到雾化后图像；然后生成随机伪影，根据伪影强度缩放后进行高斯模糊处理并添加到图像中；最后对图像应用核大小为5的高斯模糊并进行亮度和对比度的调整，以模拟术中低剂量X-ray图像的多种噪声和图像特性，生成更加接近真实手术环境的图像。在生成DRR图像时，将旋转和平移参数保存为标签作为真实变换参数。

在真实X-ray上进行训练的一大挑战是获取真实姿态参数，通常需要专家进行手动配准标注。为

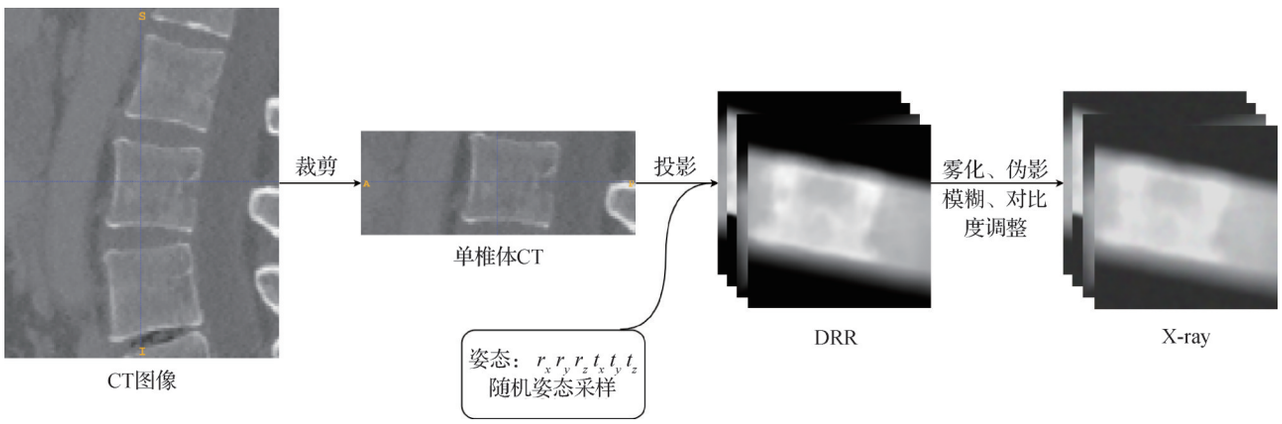

图4　数据集合成过程

Fig. 4　The process of dataset synthesis

了初步评估模型在临床应用中的性能，本文在真实X-ray上进行实验以测试所提出模型回归出姿态参数的准确性。使用的X-ray数据由极睿医疗科技有限公司所提供。

3.2　实现细节

本文的实验硬件环境为Intel i7-6850K CPU以及NVIDIA GeForce RTX 3090显卡；软件环境为Ubuntu 20.10.21操作系统，Cuda12.3。配准模型的构建和训练均由PyTorch框架实现。在训练过程中，训练模型320轮并将学习率设置为6E−5，使用Adam优化器。上述参数是根据网络的训练效率以及收敛状态确定的，训练时间大约6 h。本文网络中损失函数权重的确定是根据姿态参数与标签之间的误差值确定的，考虑到早期阶段提取的低级特征不够精确，因此辅助损失的权重设置采用由小到大的策略，逐步引导网络优化，最终权重依次设置为0.4、0.6以及0.8。

3.3　评价指标

常用的衡量配准精度的指标是平均目标配准误差（mean target registration error，mTRE），通常是通过计算配准后图像的已知标记点或特征点的位置和真实位置之间的距离获得的，具体为

$$mTRE = \frac{1}{N}\sum_{i=1}^{N}\left\| d_{\text{tar}}^{i} - d_{\text{reg}}^{i} \right\| \tag{11}$$

式中，N表示选择的点的数量，d_{tar}和d_{reg}分别表示固定图像（DRR）和浮动图像（X-ray）上的点。mTRE值越小，表示配准精度越高、图像间对应关系越精确。

此外，MAE通常用于评估预测的变参数（平移、旋转）和真实参数之间的差异。MAE越小，表明网络回归预测出的姿态误差越小，配准结果越准确。

3.4　消融实验

为验证所设计的LGFF-Reg各个组件的有效性，进行消融实验，分别从模型不同层数的设置、网络结构的设计以及损失函数超参数设置方面进行讨论。

3.4.1　不同模型层数实验

在设计深度学习模型进行单椎体图像配准时，利用双分支网络回归姿态参数，网络的层数会对配准结果产生影响，其中每层都包括下采样、特征提取以及特征融合。因此将网络的层数表示为n，作为超参数进行讨论，选取10组不同数据集进行测试，表1为n取不同值时的配准误差以及模型推理时间。

根据表1中的各项指标，本文模型层数设置为4时配准效果最佳，配准误差最小，其中mTRE和MAE的均值分别为1.38和0.009。实验结果表明，在单椎体配准任务中，浅层的模型可能因为捕捉到的信息不足，无法全面表征X-ray图像中特征与相应姿态参数之间的映射关系；而模型层数过多可能会导致特征提取的冗余，影响模型的有效性，且增加了计算开销和推理时间。

3.4.2　不同网络结构实验

单椎体配准时，X-ray通常具有有限的视野，且

表1　模型不同层数的配准效果

Table 1　Registration performance with different layer numbers in the model

模型层数	mTRE/mm	MAE	时间/s
$n=3$	1.78	0.012	**5.16**
$n=4$	**1.38**	**0.009**	5.23
$n=5$	1.60	0.010	5.51

注：加粗字体表示各列最优结果。

背景信息的存在使得从图像中提取单椎体的局部特征变得更加困难，因此需要设计一个能够有效提取单椎体特征并捕获空间映射关系的回归器结构。

本节主要验证LGFF-Reg中各个组件的有效性，包括CNN实现的局部分支（LB）、基于多头自注意力的全局分支（GB）、双分支特征融合模块（DBFF）以及辅助配准头（reg head）的设计。

第1种结构（LB），只利用CNN实现的局部分支捕获单椎体图像特征，难以获取全局信息，且背景信息会干扰CNN对椎体特征的提取，不足以提供整体的几何信息和姿态估计。

第2种结构（GB），只利用包含多头自注意力结构的全局分支捕捉椎体信息，能够更好地预测椎体位置，但在处理细粒度的局部特征时效果较差，且在处理背景噪声和无关信息时，若注意力权重分配不合理，会影响模型的配准精度。

第3种结构（LB + GB），采用双分支结构，能够提取到具有不同语义信息的特征，有助于在单椎体配准中更全面地理解和匹配椎体的结构，但在网络各层采用简单的拼接操作进行特征融合。

第4种结构（LB + GB + DBFF），利用基于注意力机制实现的模块使模型学习特征之间的依赖性和重要性，并对特征通道和位置进行加权处理，提升融合特征的表达能力、关注重要的特征部分并减少对噪声和无关信息的敏感度。

本文方法在第4种结构基础上进一步增加了辅助配准头，能够在训练过程中提供额外的监督信号、逐步细化姿态参数的映射，从而帮助模型更好地学习特征和姿态参数之间的关系。图5为LGFF-Reg在4组不同数据集的单椎体图像上的配准结果。第1行为合成的低剂量X-ray图像，第2行表示通过模型预测姿态后投影生成的配准后图像，每一列为对应不同姿态参数的不同单椎体图像。

使用不同网络结构进行单椎体配准的量化指标如表2所示。由表2可看出，当只使用局部分支或全局分支时，姿态参数预测值与真实值的误差分别为0.015和0.014，模型推理时间约为3.8 s，表明在单椎体配准任务中，仅使用堆叠的CNN编码块或Transformer的多头自注意力块难以充分提取有效的特征，网络预测值与标签之间存在较大误差。

当以简单的拼接方式融合双分支特征时，配准性能提升效果较低，因为直接拼接会导致双分支语义特征的混淆，融合后的特征不具备一致的解释性能力，还可能会导致信息冗余以及无关信息的引入，难以平衡局部和全局信息。而当增加DBFF模块后，模型所需参数量、操作更多，模型推理时间增加，但MAE降低至0.010，表明基于注意力实现的特征模块能够选择性地关注更重要的特征，做出更准确的姿态参数预测。本文方法LGFF-Reg进一步利用辅助配准头以增强监督信息，最终的配准评价指标mTRE和MAE的值分别下降到1.40 mm和0.008，且模型推理时间仅增加0.02 s，表明所设计的辅助配准头在计算量增加较小的情况下能够有效提升配准精度。

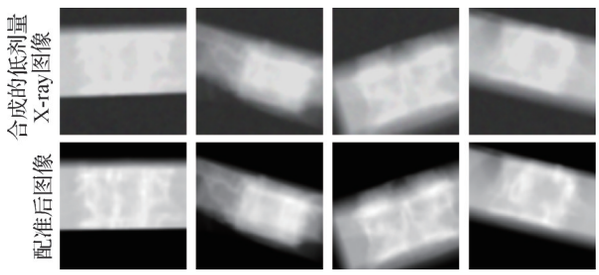

图5　不同姿态单椎体图像配准结果

Fig. 5　Registration results of single vertebra images with different poses

综上，消融实验的相关结果表明了本文设计的各个模块能够更好地获取图像特征与空间变换参数之间的映射关系，提高在单椎体图像配准任务上的性能。

表2　不同网络结构的配准评价指标

Table 2　Evaluation metrics for registration of different network architectures

网络结构	mTRE/mm	MAE	时间/s
LB	1.91	0.015	**3.83**
GB	1.72	0.014	3.86
LB + GB	1.78	0.013	3.90
LB + GB + DBFF	1.61	0.010	3.93
LGFF-Reg	**1.40**	**0.008**	3.95

注：加粗字体表示各列最优结果。

3.4.3　损失函数超参数实验

设计辅助配准头的主要目的是通过网络不同层次特征预测的姿态值与标签之间的损失引导网络中间层特征提取，因此将损失函数的权重作为超参数进行实验，选择以下4种不同策略以评估不同超参

数设置对模型性能的影响。超参数值以及配准评价指标的值如表3所示，其中，λ_1、λ_2和λ_3分别表示网络不同层次损失函数的权重。

网络最后阶段的输出是模型的预测结果，直接用于实际应用的配准任务，由于网络预测值和标签之间的差异较小，因此将最后阶段的损失函数权重固定为1，以使模型更加关注最终输出的准确性。以下是4组不同的超参数设置：

第1种策略，网络所有中间层次损失函数权重设置为1，设置网络各层对损失的贡献相同，以验证不同阶段特征提取和预测的整体效果；第2种策略，辅助损失权重较低且相等；第3种策略，辅助权重逐步递减，前期阶段损失权重较高，目的在于促进模型在早期阶段提取更有效的特征；第4种策略，辅助权重逐步递增，强调模型在后期阶段的优化效果。

由表3可看出，采用前3种超参数设置时，mTRE值约为1.6 mm，表明所采用的策略无法充分利用辅助监督信息引导模型逐步优化特征。而采用逐层递增策略时，模型性能最好，网络浅层专注于基本特征的提取，避免早期阶段与最终输出特征存在较大差异而影响模型的收敛，深层优化高层次特征，进而能够精确调整姿态，使最终回归结果更为准确。

3.5 与SOTA方法比较

为了验证本文方法的有效性，与基于迭代优化

表3 不同损失函数权重设置下的配准评价指标

Table 3 Registration evaluation metrics under different loss function weight settings

λ_1	λ_2	λ_3	mTRE/mm	MAE
1.0	1.0	1.0	1.57	0.010
0.3	0.3	0.3	1.61	0.010
0.8	0.6	0.4	1.63	0.009
0.4	0.6	0.8	**1.40**	**0.008**

注：加粗字体表示各列最优结果。

和基于深度学习的方法对单椎体2D/3D配准的精度和运行时间进行对比，选取了不同数据集下单椎体图像的配准结果进行可视化，如图6所示。其中红色、蓝色标注点分别为X-ray和配准后图像上单节椎体上的标记点，距离越近表明配准误差越小，若配准后图像未出现蓝色标注点则表示配准失败。由图6可看出，利用本文模型LGFF-Reg配准后的图像更接近相应的X-ray图像。

首先比较了3种基于优化的方法OPT-GO（optimization gradient orientation）（De Silva 等，2016）、OPT-GC（optimization gradient correlation）（De Silva 等，2016）和OPT-NGI（optimization normalized gradient information）（Otake 等，2013），定义了不同的配准指标来测量图像之间的相似性，并迭代地调整变换参数来最大化指标。这些方法需要良好的姿态初始

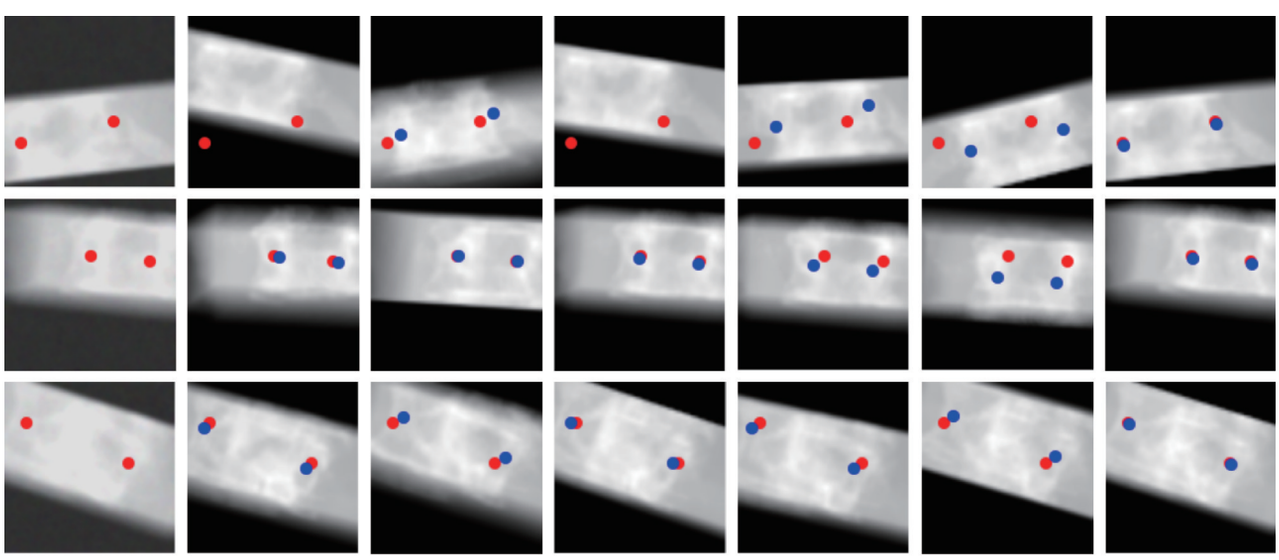

图6 不同方法单椎体图像配准结果

Fig. 6 Registration results of single vertebra images using different methods

((a) X-ray; (b) OPT-GO; (c) OPT-GC; (d) OPT-NGI; (e) ResRegNet; (f) Zhang et al., (2023); (g) LGFF-Reg(ours))

化，否则可能会陷入局部极值，导致配准失败。此外，这些方法在处理模糊和低剂量图像时缺乏鲁棒性且在临床应用中耗时较长。

在训练过程中，本文方法学习图像特征与空间变换参数之间的关系，为了验证本文模型从X-ray图像预测姿态参数时的精度，比较了两种基于深度学习的方法ResRegNet（Meng等，2022）和Zhang等人（2023）提出的方法，在训练阶段都使用2D图像作为模型输入，利用姿态标签信息监督网络训练，回归出所需6DoF姿态参数。本文方法及对比方法的mTRE、MAE和配准时间的结果如表4所示。

表4　不同方法配准图像的评价指标及时间

Table 4　Evaluation metrics and time for images registered by different methods

方法	mTRE/mm	MAE	时间/s
OPT-GO	9.74	0.049	30.55
OPT-GC	6.92	0.032	25.47
OPT-NGI	8.17	0.041	30.84
ResRegNet	9.64	0.050	**1.23**
Zhang等人(2023)	4.09	0.030	1.58
LGFF-Reg（本文）	**1.40**	**0.008**	1.38

注：加粗字体表示各列最优结果。

由表4可以看出，基于迭代优化的方法在处理低剂量X-ray图像时，计算图像相似性效率低且初始参数不合适时，会出现配准失败的情况，且耗时较高，难以满足实时性的要求。ResRegNet方法通过从CT分割切片中投影生成DRR图像，并进行阈值分割等操作后生成二值图像作为网络输入，以避免不稳定的背景干扰配准方法的鲁棒性。但对术中获取的低剂量单椎体X-ray图像存在大量噪声、伪影等且成像质量较差，ResRegNet方法无法有效地提取椎体特征，导致网络预测值偏差较大，配准后的图像与目标图像之间存在较大差距。Zhang等人（2023）的方法在提取特征过程中利用图像清晰、对称的解剖结构，但相比之下，低剂量单椎体X-ray图像特征较难识别，难以准确定位边界和结构，因此该方法用于低剂量单椎体X-ray图像配准时回归出的参数可能与真实值存在较大偏差。

相较于上述SOTA（state-of-the-art）方法，本文设计的模型通过提取图像的局部细节和全局位置信息预测姿态参数，取得了更优秀的性能。模型配准结果的mTRE值为1.40 mm，预测的姿态参数与标签之间的MAE为0.008，所需时间大约为1.38 s，表明本文方法在视野较小的单椎体X-ray图像上仍然能够实现高精度的配准效果，满足术中高精度以及实时性的要求。

3.6　临床数据集的应用

为了初步验证本文模型在真实数据集上的性能，选取了患者术中真实X-ray图像进行验证，真实X-ray以及单椎体X-ray图像如图7所示。其中，在对所需椎体进行操作时，对图像不需要的椎体所在位置进行掩膜处理。

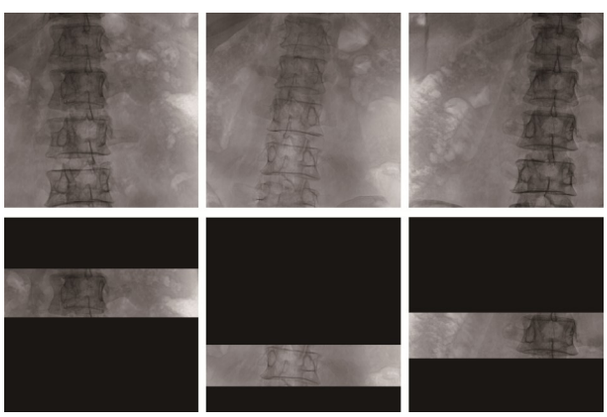

图7　真实X-ray数据集

Fig. 7　Real X-ray dataset

配准可视化结果如图8所示，其中配准前图像即网络输入图像是真实的单椎体X-ray图像，而配准后的图像是利用网络预测的姿态参数值，使用DeepDRR（Unberath等，2018）方法对相应的CT体数据进行投影后裁剪相应区域得到的DRR图像。其中红色、蓝色标注点分别为X-ray和配准后图像上单节椎体上的标记点，距离越近表明配准误差越小。由图8可看出，在真实X-ray上进行测试时脊柱位置偏差增大，这是因为DRR图像经过预处理后可能与真实X-ray图像仍存在域差异，导致在临床数据上模型性能下降。

为了解决这一问题，下一步工作将考虑收集更多的临床X-ray图像以增加真实数据的比例，利用域适应，如基于对抗学习的生成对抗网络减少合成数据集和临床数据之间的分布差异，将合成数据集上学习的特征迁移到真实数据以增强模型的性能。

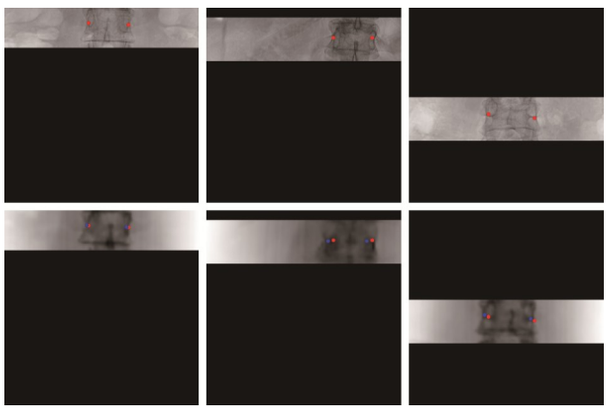

图 8　单椎体图像配准结果

Fig. 8　Registration results of single vertebra images

4　结　论

本文提出一种融合局部和全局特征的双分支网络用于脊柱的2D/3D配准，以单椎体的方式对术中X-ray图像与术前CT进行配准，解决了术中完整脊柱无法与相应CT形成刚体位置对应的问题。该方法利用卷积块和自注意力块提取单椎体图像中的结构细节以及整体位置信息，学习融合后的特征与空间变换之间的映射关系，提高回归器的性能；设计基于注意力机制实现的特征融合模块在不同尺度下自适应融合局部细节和全局上下文特征，提升模型的特征表达能力；在多尺度特征的基础上使用辅助配准头预测姿态参数增强网络的监督信息，从而进一步提升配准精度。实验表明，在单椎体图像配准任务上，与SOTA方法相比，本文模型能够取得良好的配准效果。

但本文方法也存在局限性，在对DRR图像进行雾化、模糊等操作后得到的合成DRR图像仍然与真实低剂量X-ray图像存在差异，因此在将模型迁移到真实数据上时，可能会出现配准性能下降的问题。因此，下一步工作将围绕引入真实X-ray图像进行微调或进行跨域特征迁移方面展开。

参考文献（References）

Alam F, Rahman S U, Khusro S, Ullah S and Khalil A. 2016. Evaluation of medical image registration techniques based on nature and domain of the transformation. Journal of Medical Imaging and Radiation Sciences, 47(2): 178-193 [DOI: 10.1016/j.jmir.2015.12.081]

Aubert B, Cresson T, de Guise J A and Vazquez C. 2023. X-ray to DRR images translation for efficient multiple objects similarity measures in deformable model 3D/2D registration. IEEE Transactions on Medical Imaging, 42(4): 897-909 [DOI: 10.1109/TMI.2022.3218568]

Boveiri H R, Khayami R, Javidan R and Mehdizadeh A. 2020. Medical image registration using deep neural networks: a comprehensive review. Computers and Electrical Engineering, 87: #106767 [DOI: 10.1016/j.compeleceng.20 20.106767]

Chen J Y, He Y F, Frey E C, Li Y and Du Y. 2021a. Vit-V-Net: vision transformer for unsupervised volumetric medical image registration [EB/OL]. [2024-08-27]. https://arxiv.org/pdf/2104.06468.pdf

Chen X, Diaz-Pinto A, Ravikumar N and Frangi A F. 2021b. Deep learning in medical image registration. Progress in Biomedical Engineering, 3(1): #012003 [DOI: 10.1088/2516-1091/abd37c]

Chen Z Y, Zheng Y J and Gee J C. 2024. TransMatch: a transformer-based multilevel dual-stream feature matching network for unsupervised deformable image registration. IEEE Transactions on Medical Imaging, 43(1): 15-27 [DOI: 10.1109/TMI.2023.3288136]

De Silva T, Uneri A, Ketcha M D, Reaungamornrat S, Kleinszig G, Vogt S, Aygun N, Lo S F, Wolinsky J P and Siewerdsen J H. 2016. 3D-2D image registration for target localization in spine surgery: investigation of similarity metrics providing robustness to content mismatch. Physics in Medicine and Biology, 61(8): 3009-3025 [DOI: 10.1088/ 0031-9155/61/8/3009]

Gao C, Feng A Q, Liu X T, Taylor R H, Armand M and Unberath M. 2024. A fully differentiable framework for 2D/3D registration and the projective spatial transformers. IEEE Transactions on Medical Imaging, 43(1): 275-285 [DOI: 10.1109/TMI.2023.3299588]

Gill S, Abolmaesumi P, Fichtinger G, Boisvert J, Pichora D, Borshneck D and Mousavi P. 2012. Biomechanically constrained groupwise ultrasound to CT registration of the lumbar spine. Medical Image Analysis, 16(3): 662-674 [DOI: 10.1016/j.media.2010.07.008]

Han X J, Li T T, Bai C and Yang H Y. 2024. Integrating prior knowledge into a bibranch pyramid network for medical image segmentation. Image and Vision Computing, 143: #104945 [DOI: 10.1016/j.imavis.2024.104945]

Haskins G, Kruger U and Yan P K. 2020. Deep learning in medical image registration: a survey. Machine Vision and Applications, 31(1): #8 [DOI: 10.1007/s00138-020-01060 -x]

Hille G, Saalfeld S, Serowy S and Tönnies K. 2018. Multi-segmental spine image registration supporting image-guided interventions of spinal metastases. Computers in Biology and Medicine, 102: 16-20 [DOI: 10.1016/j.compbiomed.2018.09.003]

Khameneh N B, Vazquez C, Cresson T, Lavoie F and de Guise J. 2021. Highly accurate automated patient-specific 3D bone pose and scale estimation using bi-planar pose-invariant patches in a CNN-based

3D/2D registration framework//Proceedings of the 18th IEEE International Symposium on Biomedical Imaging. Nice, France: IEEE: 681-684 [DOI: 10.1109/ISBI48211.2021.9433843]

Lan S, Li X and Guo Z H. 2023. DRT: deformable region-based transformer for nonrigid medical image registration with a constraint of orientation. IEEE Transactions on Instrumentation and Measurement, 72: #5014315 [DOI: 10.1109/TIM.2023.3273678]

Löffler M T, Sekuboyina A, Jacob A, Grau A L, Scharr A, Husseini M E, Kallweit M, Zimmer C, Baum T and Kirschke J S. 2020. A vertebral segmentation dataset with fracture grading. Radiology: Artificial Intelligence, 2(4): #e190138 [DOI: 10.1148/ryai.2020190138]

Ma M R, Xu Y B, Song L and Liu G X. 2022. Symmetric transformer-based network for unsupervised image registration. Knowledge-Based Systems, 257: #109959 [DOI: 10.1016/j. knosys. 2022. 109959]

McLaughlin R A, Hipwell J, Hawkes D J, Noble J A, Byrne J V and Cox T. 2002. A comparison of 2D-3D intensity-based registration and feature-based registration for neurointerventions//Proceedings of the 5th International Conference on Medical Image Computing and Computer-Assisted Intervention. Tokyo, Japan: Springer: 517-524 [DOI: 10.1007/3-540-45787-9_65]

Meng C, Li Y G, Xu Y Z, Li N and Xia K. 2022. A weakly supervised framework for 2D/3D vascular registration oriented to incomplete 2D blood vessels. IEEE Transactions on Medical Robotics and Bionics, 4(2): 381-390 [DOI: 10.1109/TMRB.2022.3171310]

Miao S, Piat S, Fischer P, Tuysuzoglu A, Mewes P, Mansi T and Liao R. 2018. Dilated FCN for multi-agent 2D/3D medical image registration//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI: 4694-4701 [DOI: 10.1609/aaai.v32i1.11576]

Miao S, Wang Z J and Liao R. 2016. A CNN regression approach for real-time 2D/3D registration. IEEE Transactions on Medical Imaging, 35(5): 1352-1363 [DOI: 10.1109/TMI.2016.2521800]

Mok T C W and Chung A C S. 2022. Affine medical image registration with coarse-to-fine vision transformer//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 20803-20812 [DOI: 10.1109/CVPR52688.2022.02017]

Otake Y, Wang A S, Stayman J W, Uneri A, Kleinszig G, Vogt S, Khanna A J, Gokaslan Z L and Siewerdsen J H. 2013. Robust 3D-2D image registration: application to spine interventions and vertebral labeling in the presence of anatomical deformation. Physics in Medicine and Biology, 58(23): 8535-8553 [DOI: 10.1088/0031-9155/58/23/8535]

Sanhinova M, Haouchine N, Pieper S D, Wells III W M, Balboni T A, Spektor A, Huynh M A, Guenette J P, Czajkowski B, Caplan S, Doyle P, Kang H, Hackney D B and Alkalay R N. 2024. Registration of longitudinal spine CTs for monitoring lesion growth//Proceedings of SPIE 12926, Medical Imaging 2024: Image Processing. San Diego, USA: SPIE: #129262O [DOI: 10.1117/12.3006621]

Shi J, Wang L L, Wang S S, Chen Y X, Wang Q, Wei D M, Liang S J, Peng J L, Yi J J, Liu S F, Ni D, Wang M L, Zhang D Q and Shen D G. 2020. Applications of deep learning in medical imaging: a survey. Journal of Image and Graphics, 25(10): 1953-1981 (施俊, 汪琳琳, 王珊珊, 陈艳霞, 王乾, 魏冬铭, 梁淑君, 彭佳林, 易佳锦, 刘盛锋, 倪东, 王明亮, 张道强, 沈定刚. 2020. 深度学习在医学影像中的应用综述. 中国图象图形学报, 25(10): 1953-1981) [DOI: 10.11834/jig.200255]

Sun Y X, Zhang H Q, Chen X H, Huang S D and Bai L. 2024. Fast X-ray/CT image registration based on perspective projection triangular features. Computerized Medical Imaging and Graphics, 112: #102334 [DOI: 10.1016/j.compmedimag.2024.102334]

Unberath M, Gao C, Hu Y C, Judish M, Taylor RH, Armand M and Grupp R. 2021. The impact of machine learning on 2D/3D registration for image-guided interventions: a systematic review and perspective. Frontiers in Robotics and AI, 8: #716007 [DOI: 10.3389/frobt.2021.716007]

Unberath M, Zaech J N, Lee S C, Bier B, Fotouhi J, Armand M and Navab N. 2018. DeepDRR——a catalyst for machine learning in fluoroscopy-guided procedures//Proceedings of the 21st International Conference on Medical Image Computing and Computer Assisted Intervention. Granada, Spain: Springer: 98-106 [DOI: 10.1007/978-3-030-00937-3_12]

Varnavas A, Carrell T and Penney G. 2015. Fully automated 2D-3D registration and verification. Medical Image Analysis, 26(1): 108-119 [DOI: 10.1016/j.media.2015.08.005]

Wang X Y, Zhang Z C, Xu S K, Luo X Q, Zhang B C and Wu X J. 2024. Contrastive learning based method for X-ray and CT registration under surgical equipment occlusion. Computers in Biology and Medicine, 180: #108946 [DOI: 10.1016/j. compbiomed. 2024. 108946]

Wang X Y, Zhang Z C, Xu S K, Zhang B C, Luo X Q and Hu F Y. 2024. Unsupervised cross-domain transfer network for 3D/2D registration in surgical navigation. Journal of Computer Applications, 44(9): 2911-2918 (王熙源, 张战成, 徐少康, 张宝成, 罗晓清, 胡伏原. 2024. 面向手术导航3D/2D配准的无监督跨域迁移网络. 计算机应用, 44(9): 2911-2918) [DOI: 10.11772/j.issn.1001-9081.2023091332]

Xu S K, Zhang Z C, Yao H N, Zou Z W and Zhang B C. 2023. 2D/3D spine medical image real-time registration method based on pose encoder. Journal of Computer Applications, 43(2): 589-594 (徐少康, 张战成, 姚浩男, 邹智伟, 张宝成. 2023. 基于姿态编码器的2D/3D脊椎医学图像实时配准方法. 计算机应用, 43(2): 589-594) [DOI: 10.11772/j.issn.1001-9081.2021122147]

Zhang B C, Faghihroohi S, Azampour M F, Liu S T, Ghotbi R, Schunkert H and Navab N. 2023. A patient-specific self-supervised model for automatic X-Ray/CT registration//Proceedings of the 26th International Conference on Medical Image Computing and Computer-

Assisted Intervention. Vancouver, Canada: Springer: 515-524 [DOI: 10.1007/978-3-031-43996-4_49]

Zhang Y D, Liu H Y and Hu Q. 2021. TransFuse: fusing transformers and CNNs for medical image segmentation//Proceedings of the 24th International Conference on Medical Image Computing and Computer-Assisted Intervention. Strasbourg, France: Springer: 14-24 [DOI: 10.1007/978-3-030-87193-2_2]

Zhang Z R and Chen M H. 2024. Introducing learning rate adaptation CMA-ES into rigid 2D/3D registration for robotic navigation in spine surgery [EB/OL]. [2024-08-27]. https://arxiv.org/pdf/2405.10186.pdf

Zheng G Y, Ballester M Á G, Styner M and Nolte L P. 2006. Reconstruction of patient-specific 3D bone surface from 2D calibrated fluoroscopic images and point distribution model//Proceedings of the 9th International Conference on Medical Image Computing and Computer-Assisted Intervention. Copenhagen, Denmark: Springer: 25-32 [DOI: 10.1007/11866565_4]

Zheng S Q, Yang X, Wang Y F, Ding M Y and Hou W G. 2022. Unsupervised cross-modality domain adaptation network for X-ray to CT registration. IEEE Journal of Biomedical and Health Informatics, 26(6): 2637-2647 [DOI: 10.1109/JBHI.2021.3135890]

作者简介

杨小龙，女，硕士研究生，主要研究方向为计算机视觉和医学图像配准。E-mail：yang_xiaolong2001@163.com

张战成，通信作者，男，副教授，主要研究方向为机器学习、计算机视觉和图像处理。E-mail：zczhang@usts.edu.cn

徐少康，男，硕士，主要研究方向为图像配准和医学影像。E-mail：shaokang.xu@maestrosurgical.com

张宝成，男，副主任医师，主要研究方向为创伤骨科、骨修复材料和人造血管。E-mail：baocheng255@163.com

罗晓清，女，教授，主要研究方向为多模信息融合和计算机视觉。E-mail：xqluo@jiangnan.edu.cn

胡伏原，男，教授，主要研究方向为图像处理、模式识别和人工智能。E-mail：fuyuanhu@usts.edu.cn